Bulletin UASVM, Horticulture 65(2)/2008
pISSN 1843-5254; eISSN 1843-5394

BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Alice Dorottya DOMOKOS

University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca,
3-5 Calea Manastur Street, 400372 Cluj-Napoca
dotyoka@yahoo.com

Keywords: bioinformatics, computer biology**,** sequence analysis, gene

Abstract: Bioinformatics and computational biology involve the use or development of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems usually on the molecular level. The core principle of these techniques is using computing resources in order to solve problems on scales of magnitude far too great for human discernment. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution.

INTRODUCTION

The terms bioinformatics and computational biology are often used interchangeably. However bioinformatics more properly refers to the creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. Computational biology, on the other hand, refers to hypothesis-driven investigation of a specific biological problem using computers, carried out with experimental or simulated data, with the primary goal of discovery and the advancement of biological knowledge. Put more simply, bioinformatics is concerned with the information while computational biology is concerned with the hypotheses. Bioinformatics is also often specified as an applied subfield of the more general discipline of biomedical informatics. A common thread in projects in bioinformatics and computational biology is the use of mathematical tools to extract useful information from data produced by high-throughput biological techniques such as genome sequencing. Other common problems include the study of gene regulation to perform expression profiling using data from microarrays or mass spectrometry.

MAJOR RESEARCH AREAS

Sequence analysis: In 1977, the DNA sequences of hundreds of organisms have been decoded and stored in databases. The information is analyzed to determine genes that encode polypeptides, as well as regulatory sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species. With the growing amount of data, it long ago became impractical to analyze DNA sequences manually. Today, computer programs are used to search the genome of thousands of organisms, containing billions of nucleotides. These programs would compensate for mutations (exchanged, deleted or inserted bases) in the DNA sequence, in order to identify sequences that are related, but not identical. Another aspect of bioinformatics in sequence analysis is the automatic search for genes and regulatory sequences within a genome. Not all of the nucleotides within a genome are genes. Within the genome of higher organisms, large parts of the DNA do not serve any obvious purpose. This so-called junk

DNA may, however, contain unrecognized functional elements. Bioinformatics helps to bridge the gap between genome and proteome projects--for example, in the use of DNA sequences for protein identification.

Genome annotation: In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995 by Dr. Owen White, who was part of the team that sequenced and analyzed the first genome of a free-living organism to be decoded, the bacterium *Haemophilus influenzae*. Dr. White built a software system to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

Computational evolutionary biology: Evolutionary biology is the study of the origin and descent of species, as well as their change over time. Informatics has assisted evolutionary biologists in several key ways; it has enabled researchers to:

- trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,
- more recently, compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, lateral gene transfer, and the prediction of factors important in bacterial speciation,
- build complex computational models of populations to predict the outcome of the system over time
- track and share information on an increasingly large number of species and organisms

Measuring biodiversity: Biodiversity of an ecosystem might be defined as the total genomic complement of a particular environment, from all of the species present, whether it is a biofilm in an abandoned mine, a drop of sea water, a scoop of soil, or the entire biosphere of the planet Earth. Databases are used to collect the species names, descriptions, distributions, genetic information, status and size of populations, habitat needs, and how each organism interacts with other species. Specialized software programs are used to find, visualize, and analyze the information, and most importantly, communicate it to other people. Computer simulations model such things as population dynamics, or calculate the cumulative genetic health of a breeding pool (in agriculture) or endangered population (in conservation). One very exciting potential of this field is that entire DNA sequences, or genomes of endangered species can be preserved, allowing the results of Nature's genetic experiment to be remembered *in silico*, and possibly reused in the future, even if that species is eventually lost.

Analysis of gene expression: The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed in-situ hybridization. All of these techniques are extremely noise-prone and/or subject to bias in the biological measurement, and a major research area in computational biology involves developing statistical tools to separate signal from noise in high-throughput gene expression studies. Such studies are often used to determine the genes implicated in a disorder: one might compare microarray data from cancerous epithelial cells to data from non-cancerous cells to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

Analysis of protein expression: Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample.

Bioinformatics is very much involved in making sense of protein microarray and HT MS data; the former approach faces similar problems as with microarrays targeted at mRNA, the latter involves the problem of matching large amounts of mass data against predicted masses from protein sequence databases, and the complicated statistical analysis of samples where multiple, but incomplete peptides from each protein are detected.

Prediction of protein structure: Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. (Of course, there are exceptions, such as the bovine spongiform encephalopathy - aka Mad Cow Disease - prion.) Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as one of *secondary, tertiary* and *quaternary structure*. A viable general solution to such predictions remains an open problem. As of now, most efforts have been directed towards heuristics that work most of the time.

One of the key ideas in bioinformatics is the notion of homology. In the genomic branch of bioinformatics, homology is used to predict the function of a gene: if the sequence of gene A, whose function is known, is homologous to the sequence of gene B, whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably.

One example of this is the similar protein homology between hemoglobin in humans and the hemoglobin in legumes (leghemoglobin). Both serve the same purpose of transporting oxygen in the organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes.

Other techniques for predicting protein structure include protein threading and *de novo* (from scratch) physics-based modeling.

Comparative genomics: The core of comparative genome analysis is the establishment of the correspondence between genes (orthology analysis) or other genomic features in different organisms. It is these intergenomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectra of algorithmic, statistical and mathematical techniques, ranging from exact, heuristics, fixed parameter and approximation algorithms for problems based on parsimony models to Markov Chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models.

Many of these studies are based on the homology detection and protein families computation.

Modeling biological systems: Systems biology involves the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes which

comprise metabolism, signal transduction pathways and gene regulatory networks) to both analyze and visualize the complex connections of these cellular processes. Artificial life or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms.

Protein-protein docking: In the last two decades, tens of thousands of protein three-dimensional structures have been determined by X-ray crystallography and Protein nuclear magnetic resonance spectroscopy (protein NMR). One central question for the biological scientist is whether it is practical to predict possible protein-protein interactions only based on these 3D shapes, without doing protein-protein interaction experiments. A variety of methods have been developed to tackle the Protein-protein docking problem, though it seems that there is still much place to work on in this field.

Software and Tools: Software tools for bioinformatics range from simple command-line tools, to more complex graphical programs and standalone web-services. The computational biology tool best-known among biologists is probably BLAST, an algorithm for determining the similarity of arbitrary sequences against other sequences, possibly from curated databases of protein or DNA sequences. The NCBI provides a popular web-based implementation that searches their databases. BLAST is one of a number of generally available programs for doing sequence alignment.

Web Services in Bioinformatics: SOAP and REST-based interfaces have been developed for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of the world. The main advantages lay in the end user not having to deal with software and database maintenance overheads Basic bioinformatics services are classified by the EBI into three categories: SSS (Sequence Search Services), MSA (Multiple Sequence Alignment) and BSA (Biological Sequence Analysis). The availability of these service-oriented bioinformatics resources demonstrate the applicability of web based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow management systems.

CONCLUSIONS

As a recent convert into the field, I am still amazed and excited by the beauty, complexity and challenge of analyzing information that is exploding from biological systems. I hope I have been successful to a good measure in transferring my excitement to you. I also hope that the domain of Bioinformatics and Computational Biology has given you enough exposure to help you make your own judgment about the depth and breadth of the field.

REFERENCES

1. Achuthsankar, S Nair Computational Biology & Bioinformatics - A gentle Overview, Communications of Computer Society of India, January 2007.
2. Aluru, Srinivas, 2006, Ed. Handbook of Computational Molecular Biology, Chapman & Hall/Crc,. ISBN 1584884061 (Chapman & Hall/Crc Computer and Information Science Series)
3. Baxevanis, A. D., B. F. F. Ouellette, 2005, A Practical Guide to the Analysis of Genes and Proteins, third edition, Wiley, ISBN 0-471-47878-4, eds., Bioinformatics.
4. Gilbert, D., 2004, Bioinformatics software resources, Briefings in Bioinformatics, Briefings in Bioinformatics, 2004 5(3):300-304.
5. Michael, S. Waterman, 1995, Introduction to Computational Biology: Sequences, Maps and Genomes. CRC Press. ISBN 0-412-99391-0.
6. Algorithms for Computational Biology Free MIT Course.